The BioProject Database

Connecting studies to their relevant genome records and beyond

https://www.ncbi.nlm.nih.gov/bioproject

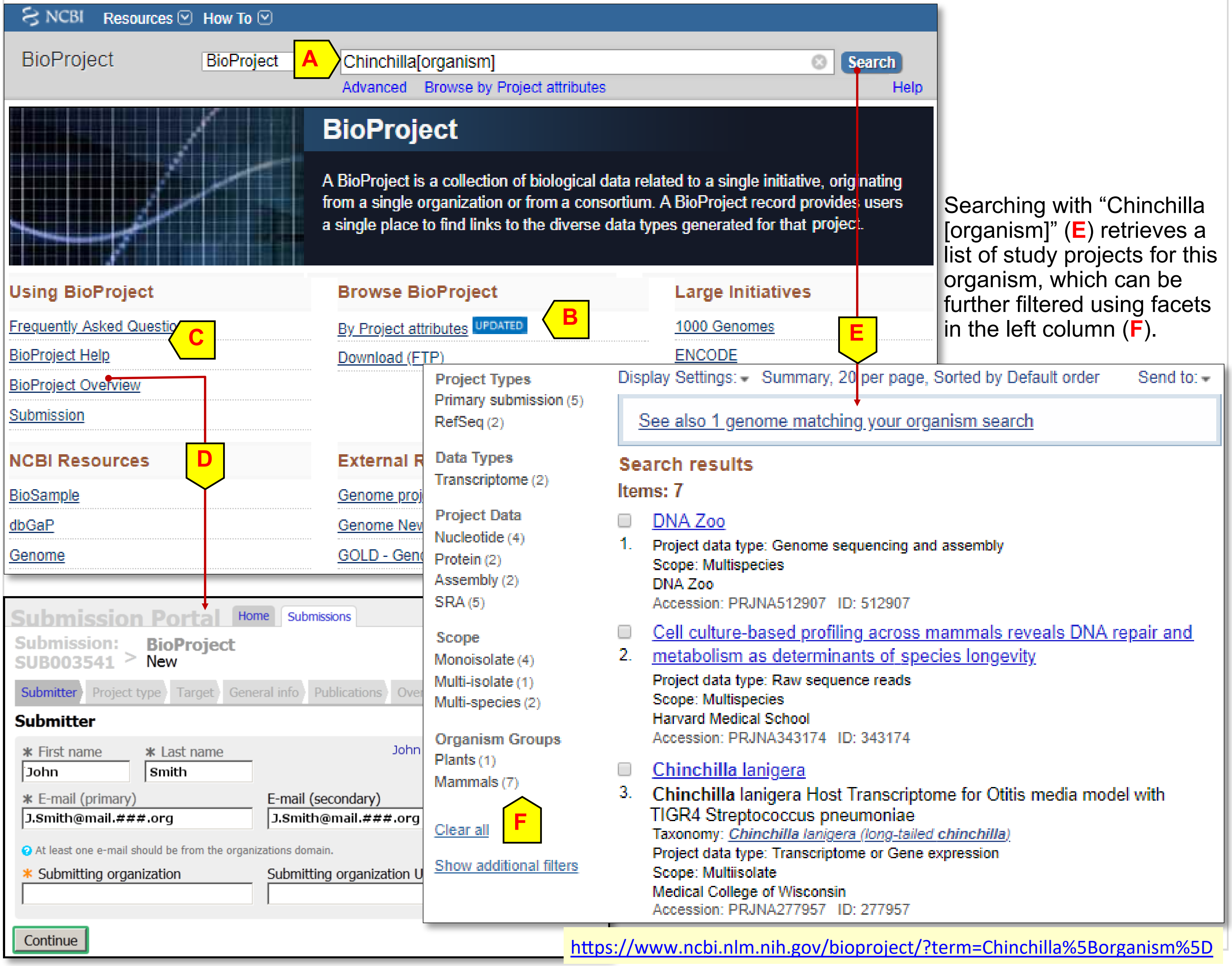
National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Scope

Biological studies in the post-genome era often produce large volumes of complex and high-throughput datasets of multiple types. For example, a single study may address topics of genomic sequencing, transcriptome, epigenome and genotype/phenotype association, or it could be further divided into multiple subprojects, each focused on a narrower field. The different characteristics of datasets generated by a complex study require that they be deposited in different databases at NCBI, such as GenBank, SRA (Sequence Read Archive), dbGaP (database of Genotype and Phenotype), Gene Expression Omnibus (GEO), or others. Registering a study in the BioProject database [1, 2] generates a central record that provides a clear overview of the scope of that study and functions as a primary key to link its divergent datasets and allow easy data access. Registration in BioProject is mandatory for new genome-scale studies with primary data submitted to the International Nucleotide Sequence Database Collaboration (INSDC). The BioProject database entries are closely related to the corresponding entries in the Genome database and can be used as a conduit to the genomic sequence data.

Accessing and registering projects

The BioProject homepage (www.ncbi.nlm.nih.gov/bioproject/) provides access to records for registered projects through a text search (**A**) or through a browser using Project attributes (**B**). Records from other NCBI databases also link to Bio-Project records. Help documentation (**C**) provides guidance for registration of projects and help in using the resource to find records. The Submission (**D**) link accesses NCBI's submission portal for the registration of new projects in BioProject. Registering a project requires credentials from an MyNCBI, NIH or eRA login.



A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.

Submission: SUB003541 > BioProject New

Submitter

John

✱ First name: John

✱ Last name: Smith

✱ E-mail (primary): J.Smith@mail.###.org

E-mail (secondary): J.Smith@mail.###.org

At least one e-mail should be from the organizations domain.

✱ Submitting organization

Submitting organization U

See also 1 genome matching your organism search

Search results

Items: 7

1. DNA Zoo
   Project data type: Genome sequencing and assembly
   Scope: Multispecies
   DNA Zoo
   Accession: PRJNA512907 ID: 512907
2. Cell culture-based profiling across mammals reveals DNA repair and metabolism as determinants of species longevity
   Project data type: Raw sequence reads
   Scope: Multispecies
   Harvard Medical School
   Accession: PRJNA343174 ID: 343174
3. Chinchilla lanigera
   Chinchilla lanigera Host Transcriptome for Otitis media model with TIGR4 Streptococcus pneumoniae
   Taxonomy: *Chinchilla lanigera (long-tailed chinchilla)*
   Project data type: Transcriptome or Gene expression
   Scope: Multiisolate
   Medical College of Wisconsin
   Accession: PRJNA277957 ID: 277957

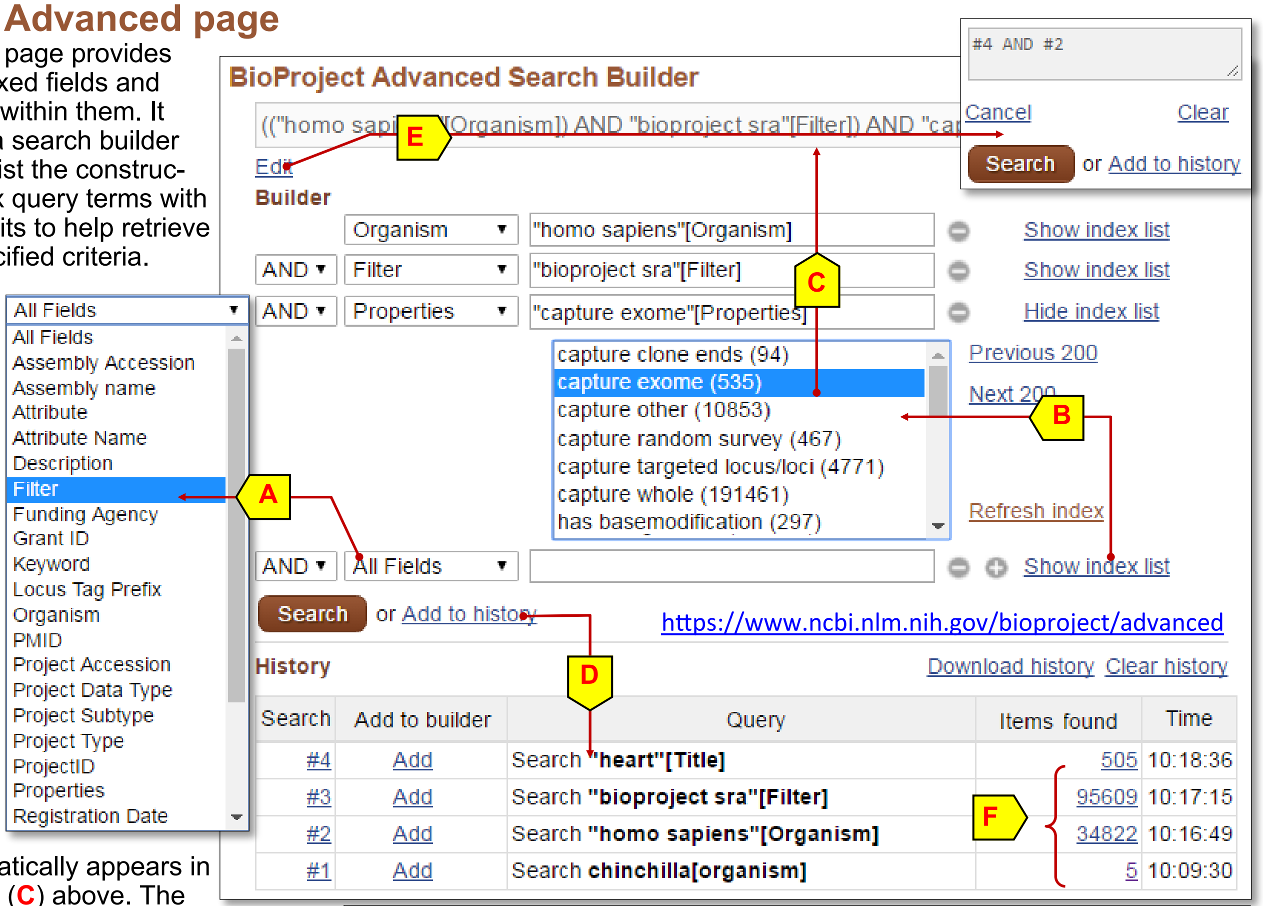
Searching with "Chinchilla [organism]" (**E**) retrieves a list of study projects for this organism, which can be further filtered using facets in the left column (**F**).

https://www.ncbi.nlm.nih.gov/bioproject/?term=Chinchilla%5Borganism%5D

Using the Advanced page

The Advanced page provides access to indexed fields and terms indexed within them. It also provides a search builder function to assist the construction of complex query terms with proper field limits to help retrieve records fit specified criteria.

In this page, clicking the index field displays available fields in a pull-down menu (**A**), which can be selected (highlighted) for use as a field limit. Clicking the “Show index list” link (**B**) adds terms indexed under the selected field. A selected term automatically appears in the search box (**C**) above. The “Add to history” link (**D**) searches the database with terms in the search box and adds the result to the history list. Unlocking the search box using the “Edit” link (**E**) allows custom input. This example combines two existing searches with AND. Clicking a number in the “Items found” (**F**) retrieves the results. The example above highlights three very important and informative fields: organism, filter, and properties.



BioProject Advanced Search Builder

((“homo sapiens”[Organism]) AND “bioproject sra”[Filter]) AND “ca

Edit

#4 AND #2

Cancel Clear

Search or Add to history

Builder

	Organism ▾	"homo sapiens"[Organism]	Show index list
AND ▾	Filter ▾	"bioproject sra"[Filter]	Show index list
AND ▾	Properties ▾	"capture exome"[Properties]	Hide index list
AND ▾	All Fields ▾		Show index list

capture clone ends (94)
capture exome (535)
capture other (10853)
capture random survey (467)
capture targeted locus/loci (4771)
capture whole (191461)
has basemodification (297)

Previous 200
Next 200
Refresh index

All Fields ▾
All Fields
Assembly Accession
Assembly name
Attribute
Attribute Name
Description
Filter
Funding Agency
Grant ID
Keyword
Locus Tag Prefix
Organism
PMID
Project Accession
Project Data Type
Project Subtype
Project Type
ProjectID
Properties
Registration Date

Search or Add to history

History

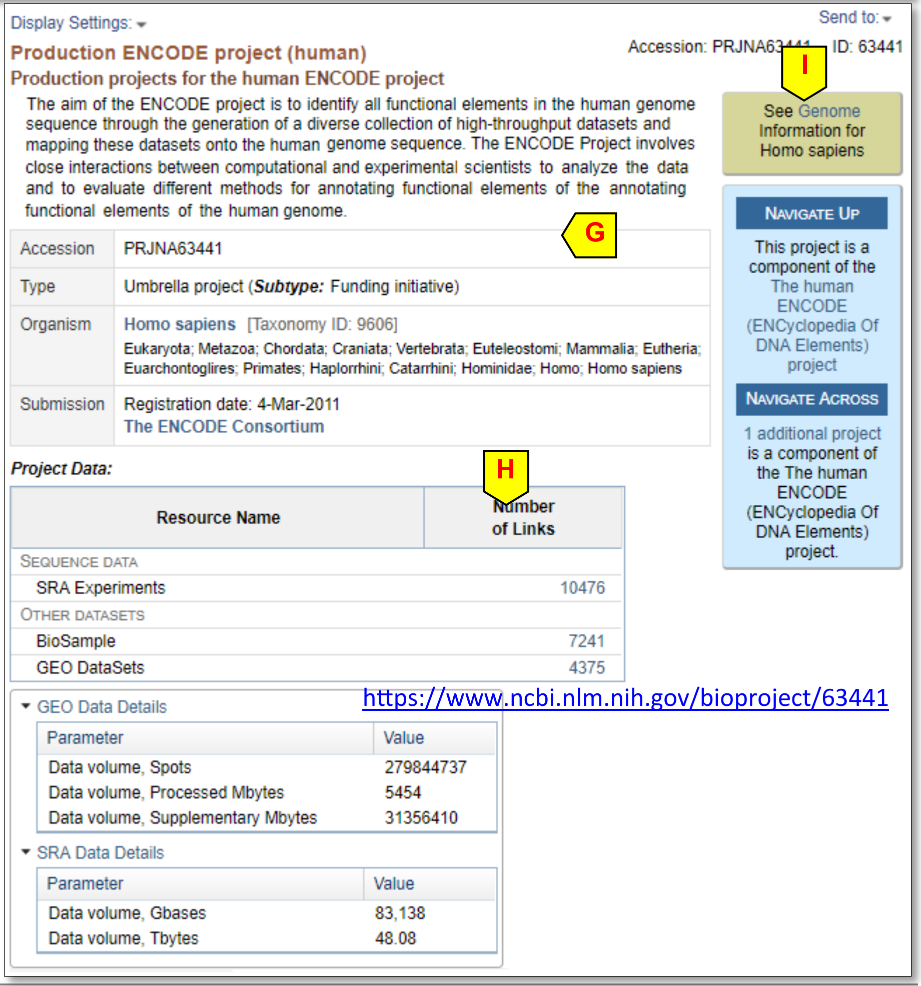
Download history Clear history

Search	Add to builder	Query	Items found	Time
#4	Add	Search **"heart"[Title]**	505	10:18:36
#3	Add	Search **"bioproject sra"[Filter]**	95609	10:17:15
#2	Add	Search **"homo sapiens"[Organism]**	34822	10:16:49
#1	Add	Search **chinchilla[organism]**	5	10:09:30

https://www.ncbi.nlm.nih.gov/bioproject/advanced

Displaying a project record

Contents displayed in project records differ for umbrella projects and data-containing primary projects. An umbrella project provides an overview serves as a centralized entry point where all sub-projects under it can be readily retrieved through links within the body of the record. The example umbrella project (shown to the right) is a human ENCODE project. The top section provides a summary on the scope of the project and the project type (**G**). The number of available links to datasets in other databases (**H**) reflects the scale of the project. Links in the “Navigate” panel (**I**) allows the navigation to the parent and sister projects.



Display Settings: ▾ Send to: ▾

Accession: PRJNA63441 ID: 63441

Production ENCODE project (human)

Production projects for the human ENCODE project

The aim of the ENCODE project is to identify all functional elements in the human genome sequence through the generation of a diverse collection of high-throughput datasets and mapping these datasets onto the human genome sequence. The ENCODE Project involves close interactions between computational and experimental scientists to analyze the data and to evaluate different methods for annotating functional elements of the annotating functional elements of the human genome.

Accession	PRJNA63441
Type	Umbrella project (***Subtype:*** Funding initiative)
Organism	Homo sapiens [Taxonomy ID: 9606] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo; Homo sapiens
Submission	Registration date: 4-Mar-2011 The ENCODE Consortium

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	10476
OTHER DATASETS	
BioSample	7241
GEO DataSets	4375

▾ GEO Data Details

Parameter	Value
Data volume, Spots	279844737
Data volume, Processed Mbytes	5454
Data volume, Supplementary Mbytes	31356410

▾ SRA Data Details

Parameter	Value
Data volume, Gbases	83,138
Data volume, Tbytes	48.08

See Genome Information for Homo sapiens

NAVIGATE UP

This project is a component of the The human ENCODE (ENCyclopedia Of DNA Elements) project

NAVIGATE ACROSS

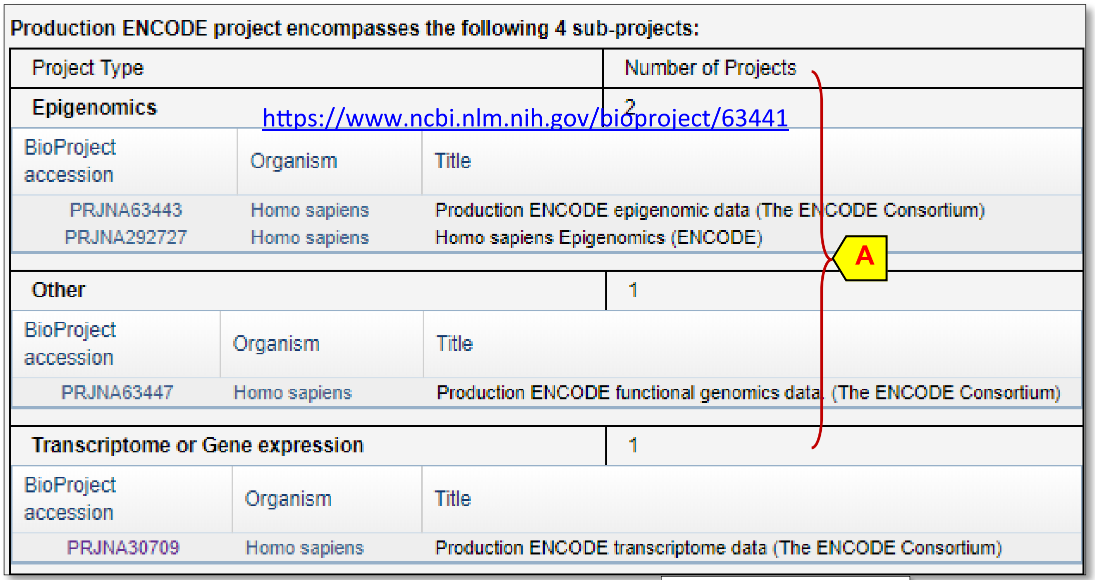
1 additional project is a component of the The human ENCODE (ENCyclopedia Of DNA Elements) project.

https://www.ncbi.nlm.nih.gov/bioproject/63441

Displaying a project record (cont.)

Actual experimental data are accessible through primary project links listed in the bottom panel (**A**).

A primary project has a simpler display and provides ***direct links*** to datasets generated by the project. In the example, the data links are shown in the "Project Data" table (**B**) as well as the more traditional Entrez links in the "Related information" section (**C**) to the right. The record also provides links to the genome record and related bioproject entries in the "Navigate" panel (**D**).



https://www.ncbi.nlm.nih.gov/bioproject/63441

Production ENCODE project encompasses the following 4 sub-projects:

Project Type	Number of Projects
Epigenomics	2

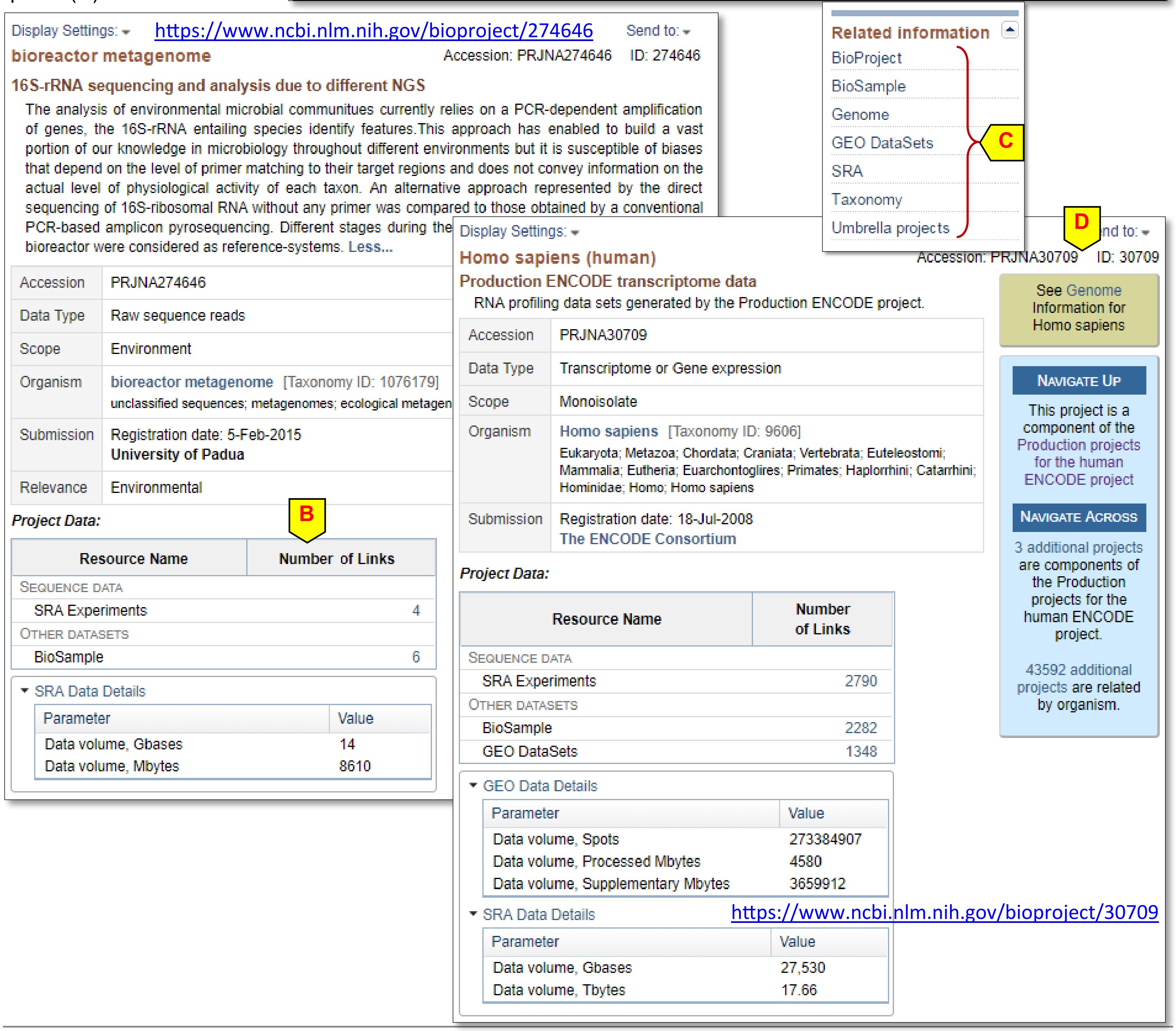
BioProject accession	Organism	Title
PRJNA63443	Homo sapiens	Production ENCODE epigenomic data (The ENCODE Consortium)
PRJNA292727	Homo sapiens	Homo sapiens Epigenomics (ENCODE)

Other	1

BioProject accession	Organism	Title
PRJNA63447	Homo sapiens	Production ENCODE functional genomics data (The ENCODE Consortium)

Transcriptome or Gene expression	1

BioProject accession	Organism	Title
PRJNA30709	Homo sapiens	Production ENCODE transcriptome data (The ENCODE Consortium)



Display Settings: ▾ https://www.ncbi.nlm.nih.gov/bioproject/274646 Send to: ▾

bioreactor metagenome

Accession: PRJNA274646 ID: 274646

16S-rRNA sequencing and analysis due to different NGS

The analysis of environmental microbial communitues currently relies on a PCR-dependent amplification of genes, the 16S-rRNA entailing species identify features.This approach has enabled to build a vast portion of our knowledge in microbiology throughout different environments but it is susceptible of biases that depend on the level of primer matching to their target regions and does not convey information on the actual level of physiological activity of each taxon. An alternative approach represented by the direct sequencing of 16S-ribosomal RNA without any primer was compared to those obtained by a conventional PCR-based amplicon pyrosequencing. Different stages during the bioreactor were considered as reference-systems. Less...

Accession	PRJNA274646
Data Type	Raw sequence reads
Scope	Environment
Organism	bioreactor metagenome [Taxonomy ID: 1076179] unclassified sequences; metagenomes; ecological metagen
Submission	Registration date: 5-Feb-2015 University of Padua
Relevance	Environmental

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	4
OTHER DATASETS	
BioSample	6

▾ SRA Data Details

Parameter	Value
Data volume, Gbases	14
Data volume, Mbytes	8610

Related information

- BioProject
- BioSample
- Genome
- GEO DataSets
- SRA
- Taxonomy
- Umbrella projects

Display Settings: ▾ Send to: ▾

Homo sapiens (human)

Accession: PRJNA30709 ID: 30709

Production ENCODE transcriptome data

RNA profiling data sets generated by the Production ENCODE project.

Accession	PRJNA30709
Data Type	Transcriptome or Gene expression
Scope	Monoisolate
Organism	Homo sapiens [Taxonomy ID: 9606] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo; Homo sapiens
Submission	Registration date: 18-Jul-2008 The ENCODE Consortium

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	2790
OTHER DATASETS	
BioSample	2282
GEO DataSets	1348

▾ GEO Data Details

Parameter	Value
Data volume, Spots	273384907
Data volume, Processed Mbytes	4580
Data volume, Supplementary Mbytes	3659912

▾ SRA Data Details https://www.ncbi.nlm.nih.gov/bioproject/30709

Parameter	Value
Data volume, Gbases	27,530
Data volume, Tbytes	17.66

See Genome Information for Homo sapiens

NAVIGATE UP

This project is a component of the Production projects for the human ENCODE project

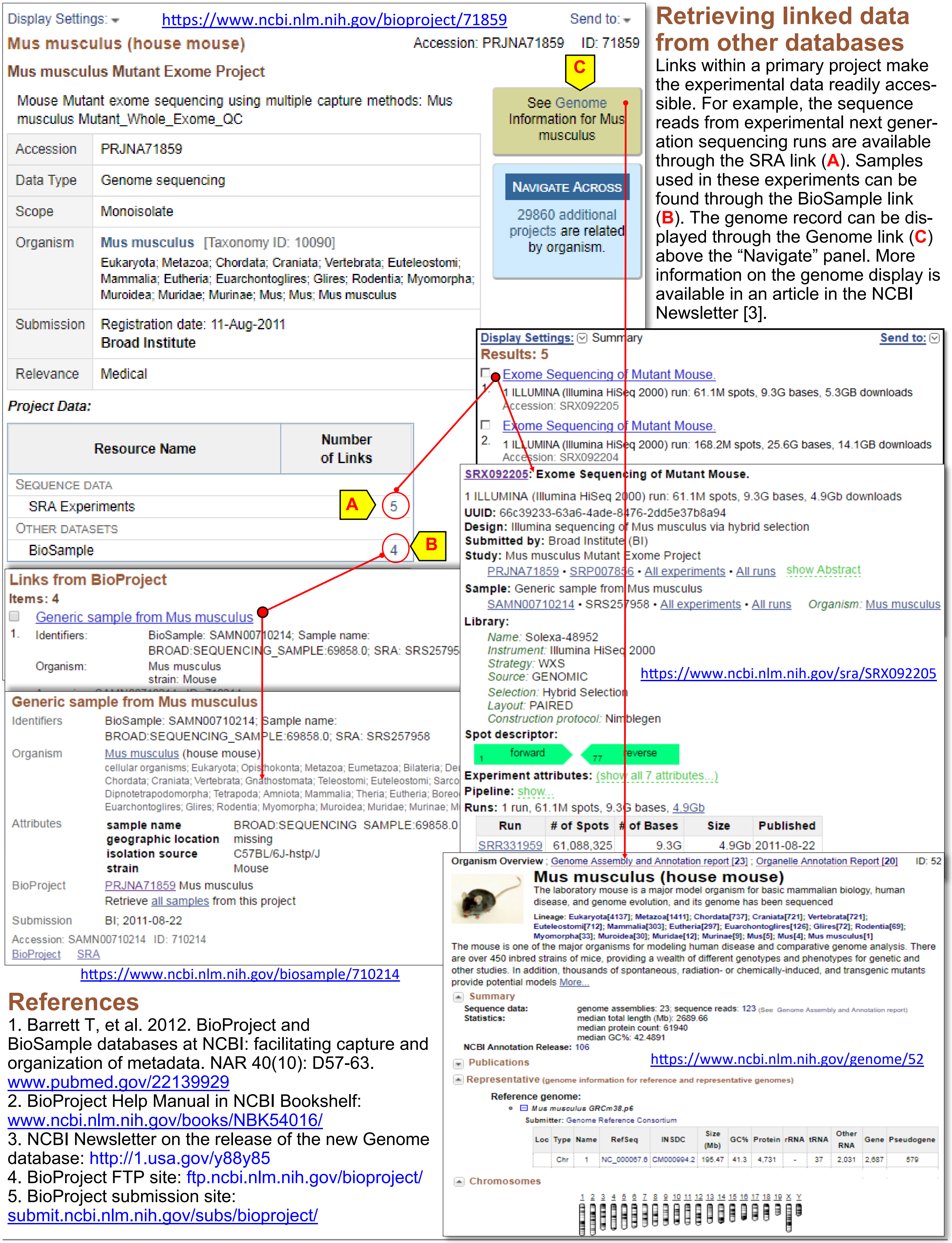
NAVIGATE ACROSS

3 additional projects are components of the Production projects for the human ENCODE project.

43592 additional projects are related by organism.

Retrieving linked data from other databases

Links within a primary project make the experimental data readily accessible. For example, the sequence reads from experimental next generation sequencing runs are available through the SRA link (**A**). Samples used in these experiments can be found through the BioSample link (**B**). The genome record can be displayed through the Genome link (**C**) above the "Navigate" panel. More information on the genome display is available in an article in the NCBI Newsletter [3].



Display Settings: ▾ https://www.ncbi.nlm.nih.gov/bioproject/71859 Send to: ▾

Mus musculus (house mouse) Accession: PRJNA71859 ID: 71859

Mus musculus Mutant Exome Project

Mouse Mutant exome sequencing using multiple capture methods: Mus musculus Mutant_Whole_Exome_QC

Accession	PRJNA71859
Data Type	Genome sequencing
Scope	Monoisolate
Organism	Mus musculus [Taxonomy ID: 10090] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae; Murinae; Mus; Mus; Mus musculus
Submission	Registration date: 11-Aug-2011 Broad Institute
Relevance	Medical

C
See Genome Information for Mus musculus
Navigate Across
29860 additional projects are related by organism.

Project Data:

Resource Name	Number of Links
Sequence data	
SRA Experiments	5
Other datasets	
BioSample	4

Display Settings: ⊙ Summary Send to: ⊙

Results: 5

□ 1. Exome Sequencing of Mutant Mouse.
1 ILLUMINA (Illumina HiSeq 2000) run: 61.1M spots, 9.3G bases, 5.3GB downloads
Accession: SRX092205

□ 2. Exome Sequencing of Mutant Mouse.
1 ILLUMINA (Illumina HiSeq 2000) run: 168.2M spots, 25.6G bases, 14.1GB downloads
Accession: SRX092204

SRX092205: Exome Sequencing of Mutant Mouse.

1 ILLUMINA (Illumina HiSeq 2000) run: 61.1M spots, 9.3G bases, 4.9Gb downloads

UUID: 66c39233-63a6-4ade-8476-2dd5e37b8a94
Design: Illumina sequencing of Mus musculus via hybrid selection
Submitted by: Broad Institute (BI)
Study: Mus musculus Mutant Exome Project
PRJNA71859 • SRP007856 • All experiments • All runs show Abstract
Sample: Generic sample from Mus musculus
SAMN00710214 • SRS257958 • All experiments • All runs *Organism:* Mus musculus
Library:
Name: Solexa-48952
Instrument: Illumina HiSeq 2000
Strategy: WXS
Source: GENOMIC
Selection: Hybrid Selection
Layout: PAIRED
Construction protocol: Nimblegen

https://www.ncbi.nlm.nih.gov/sra/SRX092205

Spot descriptor:

Experiment attributes: (show all 7 attributes...)
Pipeline: show...
Runs: 1 run, 61.1M spots, 9.3G bases, 4.9Gb

Run	# of Spots	# of Bases	Size	Published
SRR331959	61,088,325	9.3G	4.9Gb	2011-08-22

Links from BioProject
Items: 4

□ 1. Generic sample from Mus musculus
Identifiers: BioSample: SAMN00710214; Sample name: BROAD:SEQUENCING_SAMPLE:69858.0; SRA: SRS25795
Organism: Mus musculus
strain: Mouse

Generic sample from Mus musculus

Identifiers	BioSample: SAMN00710214; Sample name: BROAD:SEQUENCING_SAMPLE:69858.0; SRA: SRS257958
Organism	Mus musculus (house mouse) cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; De Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarco Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreo Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae; Murinae; M
Attributes	**sample name** BROAD:SEQUENCING SAMPLE:69858.0 **geographic location** missing **isolation source** C57BL/6J-hstp/J **strain** Mouse
BioProject	PRJNA71859 Mus musculus Retrieve all samples from this project
Submission	BI; 2011-08-22

Accession: SAMN00710214 ID: 710214
BioProject SRA

https://www.ncbi.nlm.nih.gov/biosample/710214

Organism Overview ; Genome Assembly and Annotation report [23] ; Organelle Annotation Report [20] ID: 52

Mus musculus (house mouse)

The laboratory mouse is a major model organism for basic mammalian biology, human disease, and genome evolution, and its genome has been sequenced

Lineage: Eukaryota[4137]; Metazoa[1411]; Chordata[737]; Craniata[721]; Vertebrata[721]; Euteleostomi[712]; Mammalia[303]; Eutheria[297]; Euarchontoglires[126]; Glires[72]; Rodentia[69]; Myomorpha[33]; Muroidea[30]; Muridae[12]; Murinae[9]; Mus[5]; Mus[4]; Mus musculus[1]

The mouse is one of the major organisms for modeling human disease and comparative genome analysis. There are over 450 inbred strains of mice, providing a wealth of different genotypes and phenotypes for genetic and other studies. In addition, thousands of spontaneous, radiation- or chemically-induced, and transgenic mutants provide potential models More...

▲ Summary

Sequence data: genome assemblies: 23; sequence reads: 123 (See Genome Assembly and Annotation report)
Statistics: median total length (Mb): 2689.66
median protein count: 61940
median GC%: 42.4891
NCBI Annotation Release: 106

▼ Publications

https://www.ncbi.nlm.nih.gov/genome/52

▲ Representative (genome information for reference and representative genomes)

Reference genome:
⊟ *Mus musculus GRCm38.p6*
Submitter: Genome Reference Consortium

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
	Chr	1	NC_000067.6	CM000994.2	195.47	41.3	4,731	-	37	2,031	2,687	579

▲ Chromosomes

References

1. Barrett T, et al. 2012. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. NAR 40(10): D57-63. www.pubmed.gov/22139929
2. BioProject Help Manual in NCBI Bookshelf: www.ncbi.nlm.nih.gov/books/NBK54016/
3. NCBI Newsletter on the release of the new Genome database: http://1.usa.gov/y88y85
4. BioProject FTP site: ftp.ncbi.nlm.nih.gov/bioproject/
5. BioProject submission site: submit.ncbi.nlm.nih.gov/subs/bioproject/